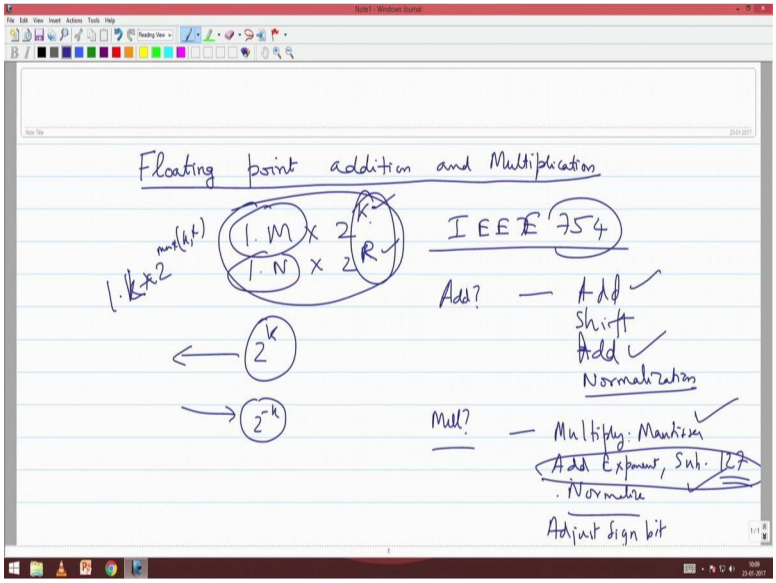
Computer Organization and Architecture
Prof. V. Kamakoti
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 08 (Part II)
Floating Point – Addition, Subtraction and Multiplication

So, what is the time required for performing floating point addition, performing floating point multiplication. Can you just tell me what would be the time required for.

(Refer Slide Time: 00:32)



Suppose I have something like 1 point M into 2 power K and 1 point N into 2 power R represented in a I triple E 754 format what it means to add and what it means to multiply this. So, what it means to add and what it means to multiply? So, how do we add 2 numbers in this format?

Student: (Refer Time: 01:32).

So, how does the; forget I triple E 754, how do you add these 2?

Student: 1 and 7.

Student: 1 and 7; the same and 1 and.

How do you make them as the same as exponent?

Student: (Refer Time: 01:43) make the lower one (Refer Time: 01:46)

Student: Make the one with less 2 power thing match the one with higher one we have to convert it (Refer Time: 01:54)

Ok.

Student: (Refer Time: 01:56).

So, how do you that yeah that is what we said by matching. So, how do you do that this is K and this is R.

Student: (Refer Time: 02:09)

Yeah. So, basically we will find out K minus R or R minus K whichever maximum and then. So, whichever is less we go and shift it by what right or left? So, whichever is less we can go and shift it by right or whichever is more we can shift it by left right. So, by then we can match these. So, what is the cost of shifting why do you shift?

Student: Multiply (Refer Time: 02:45)

So, if I shift it gets multiplied by 2 power K, if I shift K times then I it may gets multiplied by 2 power K; if I; this is shift left by shift right; K times it becomes 2 power minus K. So, we can do one of these operations.

Student: (Refer Time: 03:05).

So, shifting is take constant operation. So, I go and subtract this or add this in addition and subtraction. So, essentially after that I do a shifting. So, there is an addition involved plus there is a shifting involved. So, so that is again constant time after that what you do after had addition or shifting what do you do?

Student: Just add one (Refer Time: 03:32)

So, this gets into some value. So, you add the mantissa part. So, there is an add a shift then another add then followed by n normalisation. So, you need to normalise and bring it to someone one dot K into some 2 power of K L M 1 dot I into 2 power or max of K R R min of K R R, we have do that. So, note that a floating point addition also involves

only addition add a integer addition right. So, these 2 adds are integer adds it because I can treat them as integer and do the addition. So, how do you multiply?

Student: Sir will multiply and subtracting.

How do you?

Student: (Refer Time: 04:42) this much (Refer Time: 04:43) and add those (Refer Time: 04:45) add the.

Ok.

Student: (Refer Time: 04:50).

Add.

Student: Add exponent K plus I (Refer Time: 04:52) add the exponent and multiply the mantissa.

Multiply.

Student: M into.

Multiply mantissa; add exponent:

Student: Yes sir, then add whatever from mantissa will have will have to shift.

And normalise

Student: (Refer Time: 05:17).

That is all, is that it or something more.

Student: Sir, 1 plus 1 plus R.

Student: Is mantissa itself will considers 1 point M or.

Yeah, one point and whatever value. So, I multiply mantissa add exponent then normalise.

Student: Sign (Refer Time: 05:36).

Student: (Refer Time: 05:37).

Just sign bit that is all, did you revise.

Student: Check overflow or what (Refer Time: 05:47).

That is ok.

Student: (Refer Time: 05:49).

Well within the limit did you did you go and rewrite I triple E 754 whatever I thought that day is this correct or something wrong in this multiplication.

Student: (Refer Time: 06:01).

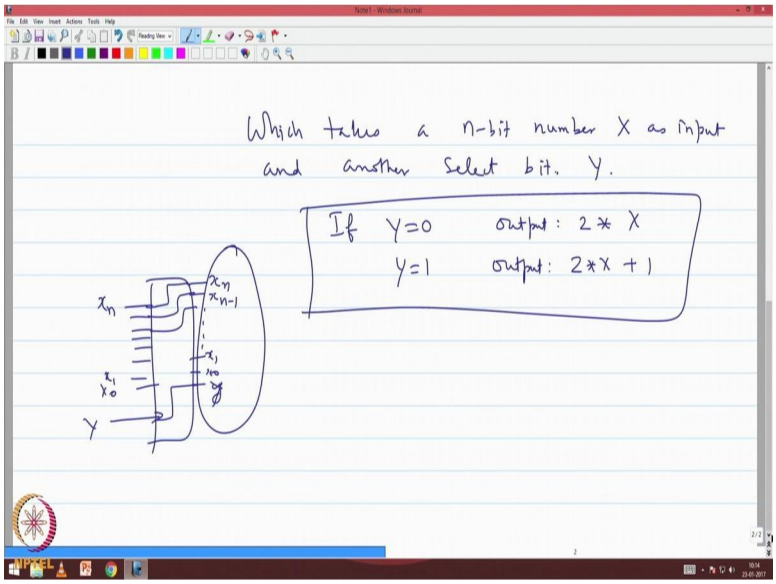
So, if I keep adding this is already excess 127 format. So, this is become excess 254 format, right. So, I have to subtract after adding exponent I have to subtract 127 format because this is excess 127 format and this is also excess 127 format if I add K plus R it becomes excess 254 for which. So, I should go and subtract. So, I should add K plus R after it is stored in IEEE 754 format and then subtract 127 format. So, this is very important crucial right. So, again please note that the multiplication also. So, this is integer multiplication this is integer addition and then normalisation.

So, you have floating point arithmetic at least addition and multiplication is as complex as integer addition or as simple as integer addition and integer multiplication it is just composing of a constant number of integer additions and or integer multiplication that is all. So, that is why even if you look at the algebraic model of computation which you use for normal algorithmic analysis you find that floating point addition and multiplication are also treated equivalent to integer addition and integer multiplication.

Because they also takes this same amount of time some constant multiple of that. So, order wise it is same. So, addition can also addition of floating point number also can be done in order log in time multiplication also in order log in time and. So, is integer addition and integer multiplication? So, this is something that we need to keep in mind. So, when we are trying to analyse an algorithm that is the theoretical analysis there is a practical analysis. So, when we do the practical analysis or even when we perform a theoretical analysis of an algorithm which we have suppose to implement and show good

speed up now these are the things that these are small that will we will talk about lot of things as you proceed in this B.tech curriculum these are some fundamental things that we should keep in mind about the practical validations of some of the models of computation that you assume while you are doing the analysis.

(Refer Slide Time: 08:42)



So, now, let me give you a very small question. So, that we can I want you to design a circuit which takes which takes a n bit number X as input and another select bit Y right if Y equal to 0 I need the output to be 2 X if Y equal to one I need the output to be 2 X plus one I will give you just 3 minutes and you design a circuit which takes a n bit number X as input and another select bit Y if Y equal to 0, I need output 2 X, if Y equal to 1 I need 2 X plus 1, how many gates; I want you to design a circuit with the we have been talking of minimisation right which you lives minimum number of gates, X is a n bit number.

Student: (Refer Time: 10:18).

What is the number of gates finally?

Student: (Refer Time: 10:21)

Ok.

Student: (Refer Time: 10:27).

Right, so how many people can understand number of gates required for this circuit is 0. So, I have X 0, X 1 till X n and this is just select bit Y the answer is just this is the M plus one. So, when Y is 0 you get 2 X just left shifted. So, you just shift it by one bit that is all right. So, there are many many interesting questions like this which will require 0 gates. So, we will discuss as we proceed in this course. So, this is one very simple example which many many placement companies ask in their interview, so then.

Student: (Refer Time: 11:45).

Ok.

Thanks.